Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia - Social and Behavioral Sciences 54 (2012) 1115 – 1124

Procedia
Social and Behavioral Sciences

EWGT 2012

15th meeting of the EURO Working Group on Transportation

Estimating traffic flow profiles according to a relative attractiveness factor

Noelia Caceres[a,*], Luis M. Romero[b], Francisco G. Benitez[c]

[a]*Senior Research Associate, Transportation Engineering, Faculty of Engineering, University of Sevilla, Sevilla, Spain*
[b]*Senior Research Associate, Transportation Engineering, AICIA, Sevilla, Spain*
[c]*Professor, Transportation Engineering, Faculty of Engineering, University of Sevilla, Sevilla, Spain*

Abstract

Traffic flow estimates play a key role for strategic and operational planning of transport networks. Although the amplitude and peak times in flows change from location to location, some consistent patterns emerge across a region. Clustering solutions appear as a powerful tool to reveal hidden trends that can easily be applied on historical traffic data to estimate traffic flows. However, these historical data traditionally are collected by detectors on only a limited number of road sections. This communication presents a methodology for estimating traffic flows using road features as clustering variables, so that it can be applied to any road section. In particular, a factor related to the attractiveness of road sections, in terms of characteristics of nearby areas, will be used to cluster road sections, deriving typical flow profiles of the resulting groups. To obtain these typical profiles, data collected by permanent detectors on a broad geographic distribution of sites across the Spanish road network have been studied. Then the flow prediction procedure for a given location is based on obtaining its attractiveness factor, finding its best match, and associating the typical flow pattern of such a group (weighted by a correction factor) to the location. The results show that the methodology make good use of historical data and, in most cases, the times of the main peaks are approximately determined. Although the prediction accuracy in the amplitude of the curves varies somewhat from location to location, the accuracy is acceptable for roads classified into groups with better similarity measurements. The applicability of the procedure to any road location makes this alternative attractive for practical applications when no detector data is available, besides no previous traffic information at the desired location is required to obtain its flow profile.

Keywords: Traffic flow estimates; clustering algorithms; attractiveness factor.

* Corresponding author. Tel.: +34 954 488135; fax: +34 954 48 73 16.
E-mail address: noeliacs@esi.us.es.

 doi:10.1016/j.sbspro.2012.09.826

1. Introduction

Traffic flow prediction is an important topic in transportation research as well as the necessary condition for the successful implementation of traffic management. Progress in data mining provides many powerful tools for effective traffic flow prediction. Among these techniques, clustering solutions exhibit great potential to identify the similarities in complex datasets and provide reliable traffic data predictions. This paper focuses on the usage of clustering methods to estimate traffic flows. Analyzing historical data by these methods, road sections can be clustered into groups according to different characteristics, with each group representing one typical flow pattern. Information about the shape of flow patterns of the resulting groups and the characteristics that are on the basis of these groups make cluster solutions appropriate for estimating traffic flows. To categorize data into different clusters, clustering procedures typically utilize characteristics of the flow as input variables, which are collected by fixed detectors (loop detectors, video cameras, radar systems, etc.). The main drawback of this methodology is that one may be interested in estimating traffic flow for road sections where there is no detector data; so that no flow characteristic is available for pattern matching. This paper proposes the usage of other road features different from those associated with the traffic flow, to cluster road locations into groups. According to the gravity model, larger places attract people more than smaller ones and places closer together have a greater attraction. Thus, a factor based on a simple form of the gravity model is used for explaining this 'geographical proximity effect' in the traffic flows on roadways. The main advantage of using this procedure is that no information about traffic flow at the desired location is required to obtain its daily flow profile. Once the groups are determined making use of historical data, the estimation procedure can be applied to any location, which is especially important when no detector data is available.

This paper is organized as follows: in Section 2, a literature review is presented; in Section 3, and after introducing the necessary concepts, the overall procedure to estimate traffic flow using geographic characteristics of road locations is presented; the estimation procedure is applied to real data and results are reported and discussed in Section 4; Section 5 concludes the paper with orientation of future work.

2. Literature Review

Clustering solutions is a powerful tool to reveal hidden trends that can easily be applied on historical traffic data. There has been much recent work on using clustering based solutions to estimate traffic information; for example, the use of K-means clustering for the prediction of motorway speed patterns (Asamer & Din, 2008) and for the segmentation of speed–density data to be applied in multiregime traffic models (Sun & Zhou, 2005), or even studies for travel time estimation using adaptive Kalman filter (Chu et al, 2005) or using artificial neural networks with clustering methods (Wei & Lee, 2003). Focusing on traffic flow predictions, different works have carried out clustering procedures to classify road sections (Weijermars & Berkum, 2005; Azimi & Zhang, 2010), most of which focused on single location data and attempted to find out the similarities of flow in different days by clustering flow series into groups. A work that applied Self-Organizing Maps to organize link flow data into relevant groups was used for pattern discovering of regional traffic status (Chen et al, 2006). The results found in this study indicated that the road links in the traffic network can be divided into several groups using different levels of feature vectors, and flows in each group have similar behaviors at the focused resolution.

The choice of clustering algorithm is an important aspect for traffic data mining. There are numerous clustering algorithms, including Bayesian clustering, hierarchical clustering and K-means clustering. A comparison using different clustering algorithms to forecast short-term freeway traffic volume was made in (Park, 2002), including a hybrid neuro-fuzzy application developed in this study. This application consisted of two components. The first one, a Fuzzy C-Means method, was for clustering the traffic flow condition and the second one, a Radial-Basis Function neural network, was for developing the estimation model associated with each of

those clusters. These results in this study were somewhat discouraging and the work concluded that better methods are needed.

Variable selection is another important issue for cluster analysis. According to the literature, clustering procedures typically utilize characteristics of flow, such as the total daily traffic flow, peak flows or times of the peak hours (Weijermars & Berkum, 2005), or even speed (Azimi & Zhang, 2010) and density data (Park, 2002), as input variables to categorize data into different groups. This paper proposes the usage of other road features different from those associated with the traffic flow, to cluster road locations into groups. The idea leads to avoid the coverage limitations using detector data for pattern matching. The main advantage of using this procedure is that no information about traffic flow at the desired location is required to obtain its daily flow profile. Hence it can be applied to a broad set of road locations.

3. Prediction methodology using clustering solutions

3.1. Introduction

This paper focuses on the usage of clustering methods to estimate traffic flow profiles by using geographic features of the roads. In particular this paper proposes a factor related to the attractiveness of road sections in terms of geographic characteristics of nearby locations. The idea aims to cluster locations (road sections) into groups using the attractiveness factor, typifying each group as one daily flow pattern using historical flow data. Thus each group is represented by its attractiveness factor centroid and its daily flow pattern. Once the groups are determined, the estimation procedure may be performed on any road location, without requiring detectors installed in such a road. For a given location, the best group is identified by matching of its attractiveness factor with any of the possible centroids. Then the typical flow pattern of such a group (weighted by a correction factor) is assigned to the location. After introducing the necessary concepts, the following subsections describe in detail the overall estimation procedure.

3.2. Clustering Method

The statistical procedure used to form groups that share similar characteristics is called cluster analysis. The goal is to organize objects into different groups or clusters, such that a group is a collection of objects “similar” to each other and are “dissimilar” to the objects belonging to other group. There are many ways to combine items into groups, an overview of clustering procedures can be found in Rui & Wunsch II, 2005. The hierarchical clustering method is the most commonly used since it avoids the need to previously specify the number of groups.

Hierarchical clustering basically forms groups by clustering cases into larger groups until all these are members of a single group. At first, all of them are considered individually, thus there are as many groups as cases. After a few iterations it reaches the final groups wanted. The criteria for deciding groups are based on either a difference or similarity matrix, where the similarity measures the closeness of cases. The different methods depend on how they estimate differences between clusters at successive steps. Among the common methods of doing this (single linkage, complete linkage, average between-groups linkage, average within-groups, centroid clustering, Ward's method...), this research has selected the average within-groups linkage. Using this method, the distance is defined as the average of the distances between all pairs of cases in the group that would result if they were combined. This tends to produce tight groups – it minimizes intra-group distances; therefore it is appropriate when the purpose of the clustering is the homogeneity within the groups. Accurate clustering requires a precise definition of the closeness between a pair of objects in multi-dimensional space, in terms of either the pair-wise similarity or distance. A variety of similarity or distance measures have been proposed and widely applied, such as Euclidean distance, cosine similarity and the Pearson correlation (Huang, 2008). In this

clustering approach, Euclidean distance has been used as the similarity measurement (*SM*) to give a numerical value to the amount of similarity between two objects (the pair-wise distances).

Finally, a way to show the progress in a hierarchical clustering procedure is provided by a dendrogram, which is a two-dimensional diagram that illustrates the fusions or divisions made at each successive stage of analysis. This graph displays the distance level at which objects and groups are joined. The optimal number of groups will be the number for which a further decrease of the number of groups leads to a high increase in variation within the groups (expressed by the rescaled distance group combine) and for which an increase of the number of groups leads to only a small decrease in variation within the groups.

3.3. Description of the Relative Attractiveness Factor

The choice of variables is another crucial issue for the cluster analysis since only those used will determine the groups. Traffic flows are the result of movements of people, so that it is seems reasonable to think that the 'geographical proximity effect' influences the number of vehicle trips on a given roadway. The gravity model is the most common formulation of the spatial interaction method, and it is therefore a popular model to use in the spatial distribution of trips (Ortuzar & Willumsen, 2001). The gravity model takes into account the size of places (as measured by population) and their distance. Since larger places attract more people than smaller ones, and closer places have a greater attraction, the gravity model incorporates these two geographic features to illustrate the macroscopic relationship between locations. By making a correlation analysis between them and flow data for a set of road locations, a strong relation is revealed. In particular, using the annual average workday traffic (*AAWT*), the results shows a proportional relation with the population ($R_{pearson}$=0.66; $R_{spearman}$=0.55), as well as an inverse correlation with the distance ($R_{pearson}$=-0.61; $R_{spearman}$=-0.84). These findings suggest that these two features can reasonably explain the abovementioned 'geographical proximity effect'. To define the attractiveness factor, a simple form of the gravity model is used: (population / distance^2). The impact of distance is squared to reflect the perception that movement is discouraged with greater impact as distance increases. There are, actually, several forms of the gravity model, but this research has considered such relationship for its simplicity and common usage. Then, the Relative Attractiveness Factor (*RAF*) of a given location is defined as the geometric mean of its attractiveness to nearby areas/cities, and it is expressed as:

$$RAF_l = \log_{10}\left(\prod_{i=1}^{n}\frac{P_i}{d_{li}^2}\right) = \log_{10}\left(geomean\left(\frac{P_i}{d_{li}^2}\right)\right) \qquad (1)$$

where i stands for each of all the possible areas near to the location l, P_i is the population size of the area i, and d_{li} is the distance between the location l and the area i.

This expression says that the force of attraction is proportional to the population size of all nearby areas divided by the square of the distance to them. The expression has utilized the geometric mean since it is a mathematical expression of the central tendency (an average) of multiple sample values. Thus the geometric mean is helpful for analyzing the effect of people concentrations because levels may vary anywhere from low to very high values. Note that the expression has taken the logarithm base 10 to compress the range of *RAF* values for a better visual representation.

Finally, it is necessary to bear in mind that the *RAF* calculation only considers the effect generated by cities that are within a defined area of influence. In particular, the calculation of the *RAF* only focuses on cities whose distance to a given road section is less than 40 km. This simplification is needed to reduce the number of areas considered in the calculation of the *RAF* for a given road section, and it is totally coherent with the assumption that nearest cities are responsible for most the traffic supported by a given road.

3.4. Estimation of traffic flow by matching the Relative Attractiveness Factor

This research aims to cluster locations (road sections) into groups using the relative attractiveness factor, typifying each group as one daily flow pattern using historical flow data. The historical data used in this study were provided by permanent loop detectors (24 hours a day, 365 days a year) installed on 374 road sections with different traffic background and characteristics in the Spanish road network (DGT, 2008). The dataset included traffic flow data in-terms of annual average workday traffic (*AAWT*). This study has used annual average measures to just consider trends introduced by population (usual residents plus visitors) which are difficult to disaggregate on a continuous basis. For each type of day, daily flow profiles have been constructed for each road by combining the *AAWT* measurements. However, this study only focuses on working days. Once the groups are determined, the flow prediction procedure for a given location is fundamentally based on obtaining its *RAF*, finding its best group, and associating the typical flow pattern of such a group (weighted by a correction factor) to the location. These steps are described in detail.

STEP 1. Relative Attractiveness Factor Calculation

As already mentioned in section 3.3, the evaluation of the *RAF* only focuses on cities within the influence area of the road location (less than 40 km). This step first requires to identify which are those cities in order to point out the population and the distance needed for the *RAF* calculation. Then, the *RAF* of all locations in the historical dataset will be calculated based on the expression (1).

STEP 2. Clustering Procedure

The purpose of the clustering analysis is to place objects into groups or clusters according to a clustering variable, in such a way that objects in a given group tend to be similar to each other in some sense, and objects in different groups tend to be dissimilar. To conduct this research, the hierarchical clustering is the procedure chosen, using the average within-groups linkage to estimate differences between groups at successive steps, and the Euclidean distance as the measure of dissimilarity between two vectors. Based on a dendrogram analysis, an eight-group solution is chosen to cluster road sections according to the *RAF* dataset. Each resulting group is represented by one attractiveness factor centroid.

STEP 3. Typical Flow Pattern Discovering

Once the road locations are clustered into groups using their *RAF* data, each group is typified as one daily flow pattern using historical flow data. The typical flow patterns for groups are determined as the average of the flow profiles of all road sections within each group. Fig 1 shows the distribution of traffic by time-of-day (in percentage of daily traffic) for the patterns obtained using the historical dataset. The figure also indicates the number of road sections in each group, *N*, and the annual average workday traffic (*AAWT*) in order to specify the traffic load on road sections within each group during a typical working day. They reflect the behavior of users in the region under study with the peak periods closely related to the regular daily routine (the start of working hours, lunch time...). Some of these patterns consist of the classic two peaks associated with commuter trips; these are the am-peak period (07:00–09:00 hours) and the pm-peak period (16:00–18:00 hours), when the majority of people travel to, and from, work or school. In other patterns, traffic flow generally starts increasing between 5:00 and 6:00 am, and are maintained (or even continue increasing) until sometime in the afternoon. The specific amplitude of the pattern curves varies somewhat from location to location, depending on the volume of traffic on a road (particularly in relation to roadway capacity). During the night, between 24:00–04:00 hours, traffic flows are at the lowest for all patterns, when the majority of people are resting. In Figure 1, every typical flow pattern has its own color (gray scale), while the profiles of road sections within a group are in black color. This makes it easy to distinguish the similarities and differences between the typical flow pattern of a group and the daily flow profiles of road sections within it. Notice that the variation of flow profiles of road sections within each resulting group is quite large for some groups. This is due to the clustering procedure does not lead to find homogeneity within the groups in terms of flow profiles but according to *RAF* values.

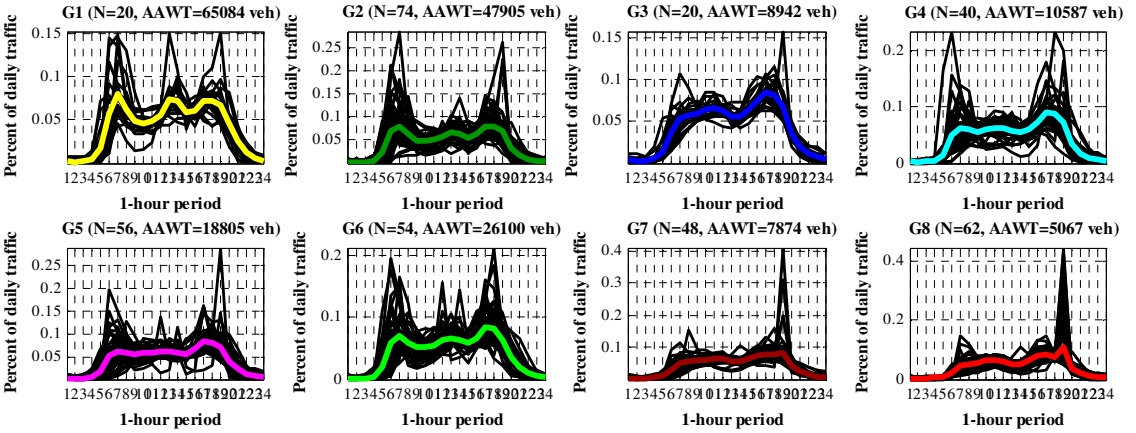

Fig. 1. Typical Patterns for the distribution of traffic by time-of-day (in percent of daily traffic) for each group.

Once the centroids of groups (*RAF* values) and the corresponding traffic flow patterns are established, the procedure for traffic flow estimation at one location only requires to carry out step 4, which is described below. It is necessary to highlight that the number of groups depends on the characteristics of the studied region and, in this case, an eight-cluster solution has been selected. The corresponding clustering results (*RAF* centroids and groups) can be considered valid for a certain time (eg: 6 months, 1 year, etc.) as long as the characteristics of the sample are more or less stable. But it would be desirable to perform periodically a new clustering stage (step 1-2-3) using updated data, in order to remodel the underlying changes in individual behavior and travel patterns.

STEP 4. Traffic Flow Estimation

Once each resulting group is represented by its attractiveness factor centroid and its flow pattern, a daily flow profile for a given location will be estimated by applying the following sub-steps:

- STEP 4.1. Relative Attractiveness Factor Evaluation for a new location

The RAF of the new road location is calculated.

- STEP 4.2. Group identification

For the new location, the best group is identified by matching its *RAF* with any of the possible *RAF* values of the centroids, using an Euclidean metric.

- STEP 4.3.Correction Factor Evaluation

After identifying the group into which the location is classified, the estimation procedure assigns the *RAF* centroid and the typical flow pattern of such a group to the location. However, this pattern cannot be directly regarded as the estimated profile since it is an average pattern for all road sections belonging to such group, and the variability in flow profiles within a group may be large, as can be seen in Fig. 1. Thus a correction factor (*CF*) will be applied to that corresponding typical flow pattern. This factor is designed to correct the differences in matching during the group identification. It is defined as the quotient between the *RAF* value of a given location k and the *RAF* value of the centroid of the associated group:

$$CF_k = \frac{RAF_k}{RAF_{Gi}} \quad \text{where } Gi \text{ stands for the group into which the location } k \text{ has been clustered} \qquad (2)$$

- STEP 4.4.Estimation of the Daily Flow Profile

The estimate of the daily flow profile for the location will be accomplished by weighting the typical flow pattern of the linked group by the corresponding correction factor. The main advantage of this procedure is that it can be applied to any location, without requiring detectors installed in such a road.

4. Experimental results

This paper has proposed a procedure to estimate daily flow profiles using a clustering approach based on geographic features of the roads. The main advantage of using this procedure is that no traffic information at the desired location is required to obtain its flow profile. Once the groups are determined making use of historical dataset, the estimation procedure can be applied to any road location by just calculating its relative attractiveness factor. This section shows the experiment results after applying the procedure described in Section 3 to another dataset from different road locations. In particular, a total of 94 road locations compounded the testing dataset. On one hand, the population and distance to cities within the influence area of these locations was determined in order to obtain the *RAF* values and to execute the estimation procedure. On the other hand, traffic flow information collected by loop detectors were also available for these location, in-terms of *AAWT*. Later a daily flow profile was constructed for each location by using these *AAWT* measurements. Thus, the obtained flow profile is used as the real one to compare with the estimate. Notice that none of the roads in the testing dataset is included into the historical data set used for building the typical flow patterns.

4.1. Error indices

Mean Absolute Relative Error

For the analysis of the estimation error, the Mean Absolute Relative Error (*MARE*) between the estimated value and the observed one at the location k is used:

$$MARE_k[\%] = \frac{1}{H} \sum_{h=6:00h}^{21:00h} \frac{abs\left(y_k^{pred}(h) - y_k^{obs}(h) \right)}{y_k^{obs}(h)} \times 100 \qquad (3)$$

where H is the number of observed 1-hour periods, $y_k^{obs}(h)$ is an observed value in the 1-hour period h at the location k, and $y_k^{pred}(h)$ is a estimated value in the 1-hour period h at the location k. This index is expressed in terms of percentage. It is necessary to highlight that this mean error rate has been calculated on the basis of the time period between 06:00 and 21:00 hours. For all typical flow patterns, traffic flows start to become significant at 6h and are maintained until sometime in the afternoon, while the number of vehicles drops substantially late at night. Thus the traffic flow is of interest for practical applications in the time period between 6:00 and 21:00h, hence the error analysis has been focused on this time period.

Clustering Similarity Measurement Coefficient

Another index to evaluate the prediction quality is the Clustering Similarity Measurement Coefficient (*CSMC*). As mentioned in Section 3.4, the variation of flow profiles for road sections within each resulting group is large because the clustering procedure does not reach homogeneity within the groups in terms of traffic flow profiles but according to *RAF* values. The greater the homogeneity (similarity) within the group in terms of flow profile is, the better the fitting of its typical pattern of daily flow profile is. To examine whether the typical pattern of each group is well-matched to all road sections within it, a clustering similarity measurement coefficient is defined. This coefficient analyzes the variation within the groups in terms of flow profiles. During the clustering procedure by *RAF* data, the similarity was evaluated according to the Euclidean distance metric. However, the mentioned coefficient cannot be taken as a similarity measurement because the order of magnitude at flow level for a group varies according to the mean amplitude of the pattern curves. In order to maintain the scale ratio among groups, the used similarity measurement coefficient should be also based on Euclidean distance but normalized by the module of the typical pattern of the group. Thus, the *CSMC* for a given group *Gi* is defined as follow:

$$CSMC_{Gi} = \frac{1}{N_{Gi}} \sum_{\mathbf{p} \in Gi} \frac{\sqrt{\sum_{h=1}^{24} \left(p_h - g_h^{Gi} \right)^2}}{\left\| \mathbf{g}^{Gi} \right\|} \quad (4)$$

where **p** is an object (daily flow profile) within the group *Gi*; $\mathbf{g}^{Gi}$ is the typical flow pattern of the group *Gi*; p_h, g_h^{Gi} are the flows in the 1-hour period *h*; and N_{Gi} is the number of objects within the group *Gi*.

Now, we have at our disposal a normalized similarity measurement to quantify how similar flow profiles within a group are, which gives the magnitude of difference between the profiles within a group. This measurement is closely related to the error margin in the estimated flow profile for a given location against its real flow profile. The error margin for flow profile estimates using the groups that have better (lower) similarity measurements will be smaller than those estimates using other groups.

4.2. Results

In order to carry out a comprehensive analysis, first the results after applying the estimation procedure (step 4) at a particular location are presented. Next, the results obtained for all road locations that compound the testing dataset are analyzed. For the particular road location *k* (E-272-0, PK 57.65), the *RAF* is obtained taking into account its influence area (step 4.1), resulting into a *RAF* value of 3.22. By matching using the Euclidean metric, this *RAF* value results closer to the centroid of group 6; so this is the best group (step 4.2). Taking into account the *RAF* value of this location and the *RAF* centroid of group 6, the correction factor is calculated (step 4.3). Finally, this correction factor is applied to the typical flow pattern of group 6 (step 4.4), resulting in the daily flow profile showed in Fig. 2 (blue/continuous line). To make a comparison, this figure also displays the observed daily flow profile at the location (red/dashed line), revealing that the estimates follow the peaks and valleys of the observed flow curve within low error level (*MARE*=12.82%). This error rate can be regarded as admissible considering that commonly-used sensors like loop detectors are also subject to errors. First, they tend to undercount vehicles; second, detectors tend to count vehicles in neighboring lanes in addition. The standards defined are that the total traffic volume should not vary from reality by more than 20% (Lehnhoff, 2004). Then, the error levels obtained are within the limit for fulfilling the standards. For this road location *k*, the values of the absolute relative error (*ARE*) for the most of the 1-hour periods between the 6:00 and 21:00 hours are smaller than the limit of 20%, being the *MARE*=12.82%. These error levels reveal that, in this case, the typical flow pattern of the group into which the location has been classified properly fits to the daily flow profile of the location. Hence this estimate can be regarded as suitable to be used in practical applications when neither detector nor flow information is available for this location.

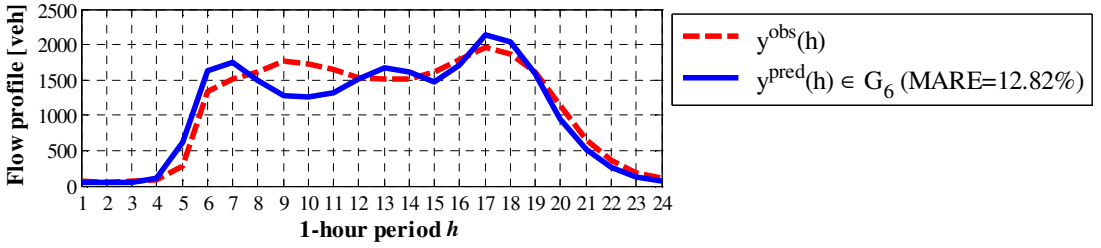

Fig. 2. Vehicle flows observed and estimated in each 1-hour period at location *k* (belonging to Group 6).

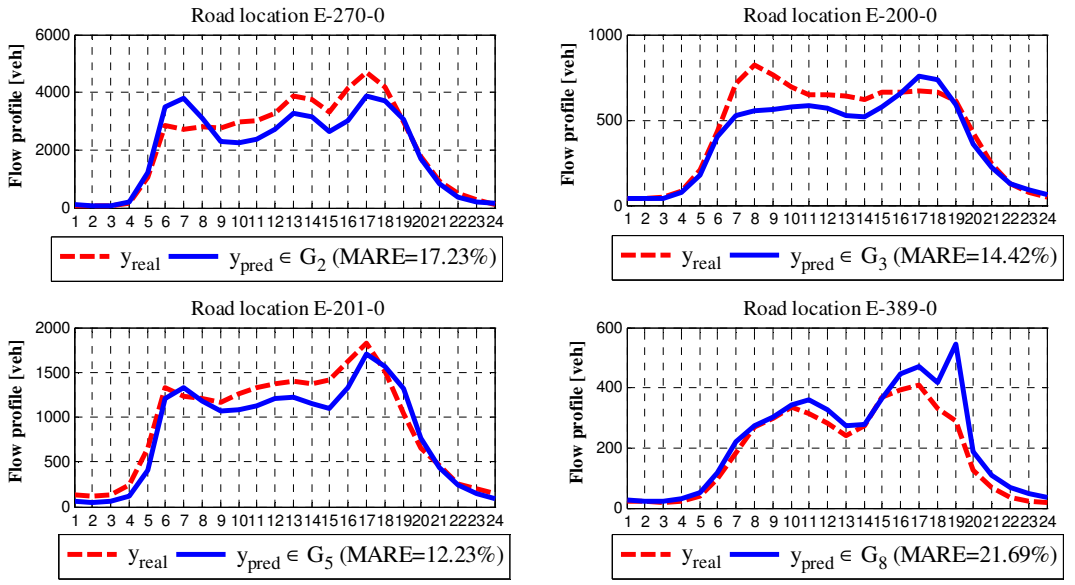

Fig. 3 Vehicle flows observed and estimated in each 1-hour period at different locations.

The same procedure is applied to the rest of road locations in the testing dataset. Figure 3 displays flow estimates for some locations together with their values of *MARE*. For most locations, the time of the main peaks is approximately determined, although the prediction accuracy in the amplitude of the curves varies somewhat from location to location. These error levels depend on whether the typical flow pattern of the group is well-matched for daily flow profiles for road sections belonging to such a group. Notice that, after identifying the group into which a given location is classified, the estimation procedure assigns the typical flow pattern of this group to the location (weighted by a correction factor). Thus the estimation error increases for locations classified into groups that have high variability in flow profiles. An indicator of pattern representativeness (or similarity) is the *CSMC* defined in Section 4.1. Table 1 lists this coefficient for each group, showing that groups 1, 3, 4, 5 and 6 are those whose profiles are more similar within the group, which is also consistent with the visual evaluation of Fig. 1. The analysis of the *MARE* rate obtained for the locations, according to the group into which it has been classified, reveals that the *CSMC* is a suitable indicator of the estimation quality. The comparative, showed in Table 1, illustrates that, for almost all groups with low values of *CSMC* (groups 1, 3, 4, 5 and 6), most of the estimates result in *MARE* rates smaller than 20%. So the errors are within the limit of 20% for fulfilling the standards. These results can be regarded as reasonable taking into account that no previous traffic information has been used as input data to obtain its flow profile.

Table 1. *CSMC* and percentage of road locations within each group with MARE<20%

	G1	G2	G3	G4	G5	G6	G7	G8
CSMC	0.029	0.088	0.038	0.054	0.056	0.051	0.058	0.101
[%] of locations with *MARE<20%*	51%	22%	50%	49%	71%	51%	42%	15%

5. CONCLUSIONS AND FUTURE DEVELOPMENTS

This communication has proposed a methodology for estimating traffic flows using a clustering approach based on geographic features of the roads (population and distance to nearby cities). Its main advantage is that no traffic information at a given location is required to obtain its flow profile. According to the gravity model, larger places attract people more than smaller ones, and places closer together show greater attraction. Thus, a Relative Attractiveness Factor (*RAF*) based on a simple form of the gravity model is used as clustering variable for explaining this 'geographical proximity effect' in traffic flows on roadways. Once the groups are determined making use of historical flow dataset, the estimation procedure can be applied to any road location by just getting its *RAF* ratio. The experiment results show that, for most of the studied locations, the time of the main peaks is approximately determined. Although the estimation accuracy in the amplitude of the curves varies somewhat from location to location, the accuracy is acceptable for roads classified into groups with better similarity measurements. For these cases, reasonable values of *MARE* are obtained, and hence these flow estimates can be regarded as suitable to be used in practical applications when no traffic information is available. Further research can be aimed at determining how primary and secondary roads affect traffic patterns, incorporating the road type (particularly in relation to roadway capacity) into the clustering process.

Acknowledgements

This research was financed by the Spanish Ministries of Public Works and Science through public-private cooperation programmes related to transport and infrastructures (SIMETRIA, REF P63/08). One of the authors, L.M. Romero, acknowledges the support of the grant PTQ-11-04952 from the Spanish Ministry of Science through the INNCORPORA-PTQ Programme.

Appendix A. References

Asamer, J., & Din, K. (2008). Prediction of Velocities on Motorways by k-Means Clustering. Proc. of the 7th Mexican International Conference on Artificial Intelligence (MICAI), pp. 399-403.

Azimi, M., & Zhang, Y. (2010). Categorizing Freeway Flow Conditions by Using Clustering Methods. In TRR: Journal of the Transportation Research Board, No 2173, Washington, D.C., pp. 105-114.

Chen Y.D., Y. Zhang, & Hu, J.M. (2006). Pattern discovering of regional traffic status with self-organizing maps. Proc. of the 9th Int. Conf. on ITS (ITSC). Toronto, Canada, pp. 647-652.

Chu, L., Oh, J-S., & Recker, W. (2005). Adaptive Kalman Filter Based Freeway Travel Time Estimation. Transportation Research Board 84th Annual Meeting, CD-ROM, paper no. 1118.

DGT, General Directorate of Roads (2008). Traffic Map 2008. Spanish Ministry of Public Works.

Huang, A. (2008). Similarity measures for text document clustering. Proc. of the Sixth New Zealand Computer Science Research Student Conference (NZCSRSC), Christchurch, New Zealand, pp. 49–56.

Lehnhoff, N. (2004). Quality of automatic data collection with loop detectors, in Proc. 2nd Int. Symp. Netw. Mobility, Germany.

Ortuzar, J.de D., & Willumsen, L.G. (2001). Modelling Transport. 3rd Edition. Wiley and Sons.

Park, B. (2002). Hybrid Neuro-Fuzzy Application in Short-Term Freeway Traffic Volume Forecasting. In TRR: Journal of Transportation Research Board, No. 1802, Washington, D.C., pp. 190-196.

Rui, X., & Wunsch II, D. (2005). Survey of clustering algorithms. IEEE Trans. on Neural Networks, vol. 16, no. 3, pp. 645-677.

Sun, L., & Zhou, J. (2005). Development of Multiregime Speed–Density Relationships by Cluster Analysis. In TRR: Journal of Transportation Research Board, No. 1934, Washington, D.C., pp. 64-71.

Wei, C.H., & Lee, Y. (2003). Development of freeway travel time forecasting models using artificial neural networks. 10th World Congress on Intelligent Transportation Systems, Madrid, CD-ROM Paper.

Weijermars, W., & van Berkum, E. (2005). Analyzing highway flow patterns using cluster analysis. 8th International IEEE Conference on Intelligent Transportation Systems (ITSC), Vienna, Austria.